Incorporating Human Dimension in Autonomous Decision-Making on Moral and Ethical Issues

Bipin Indurkhya, Joanna Misztal-Radecka
Jagiellonian University, Cracow (Poland)
bipin.indurkhya@uj.edu.pl, asiamsztl@gmail.com

Abstract

As autonomous systems are becoming more and more pervasive, they often have to make decisions concerning moral and ethical values. There are many approaches to incorporating moral values in autonomous decision-making that are based on some sort of logical deduction. However, we argue here, in order for decision-making to seem persuasive to humans, it needs to reflect human values and judgments. Employing some insights from our ongoing researchusing features of the blackboard architecture for a context-aware recommender system, and a legal decision-making system that incorporates supra-legal aspects, we aim to exploreif this architecture can also be adapted to implement a moral decision-making system that generates rationales that are persuasive to humans. Our vision is that such a system can be used as an advisory system to consider a situation from different moral perspectives, and generateethical pros and cons of taking a particular course of action in a given context.

Introduction: Human dimension in moral and ethical issues

From Plato's *Republic* through Kant's *Groundwork for the metaphysics of morals,* the traditional view of morality has been that moral values are normative: objective, rational and mind-independent. However, in recent years, this view has been challenged on many fronts: neurological evidence has shown that emotional attitudes form a cornerstone of moral reasoning (Damasio 2005; LeDoux 2003); psychologists have articulated principles of moral reasoning based on how people actually make such decisions in different contexts (Ariely 2010, 2013; Bucciarelli, Khemlani & Johnson-Laird 2008); and philosophers have incorporated these findings in their theories (Churchland 2012; Herman 2011, 2015; Johnson 2014).

On the other hand, robots and autonomous systems are becoming a part of our everyday lives at an alarming rate, and are assuming more and more decision-making roles. Some of these roles involve making moral choices, and our society is starting to take a serious look at them. Companionship, healthcare, and war are three areas where ethical and moral issues are lagging behind the pace of technology, though they are certainly being hotly debated (Arkin, Ulam & Wagner 2012; Levy 2008; Lin, Abney &Bekey 2014).

In keeping with this contemporary perspective on moral reasoning, we aim to design a system to generate moral arguments that are persuasive for humans. The importance of incorporating human dimension in decision-making is perhaps best illustrated by a key problem facing the autonomous driverless cars: they follow the traffic rules, avoid obstacles, and maneuver through traffic adroitly, but are not able to anticipate the behavior of human drivers who do not always follow the traffic rules (Richtel & Dougherty 2015). In fiction, this aspect of moral decision-making is illustrated in the films *I, Robot* and *Sophie's Choice.* For example, in *I, Robot,* the protagonist, Del Spooner, distrusts robots because a robot saved his life after a car crashbased on the probability estimates, leaving a young girl to die.

Our first objective is to implement an advisory system that examines the morality of taking an action in a given situation considering different ethical considerations. If we treat possible ethical decisions as the set of items that may be selected by the user, this problem may be seen as akin to designing a recommendation system. Hence we would like to use our ideas and experience in designing a Context-Aware Recommendation system, CARE (Misztal& Indurkhya 2015). A key feature in the design of

CARE is to have an explanation accompanying each recommendation. This feature becomes even more important for a morality-advising system, as our goal is to generate, for each course of action, a moral justification that is intuitively persuasive for humans.

This paper is organized as follows. In the next section, we present some examples to illustrate how different moral justifications can be made for different courses of actions. In the following section, we propose the architecture fora morality-advising system. Finally, we conclude the main points of this paper and point out future research directions.

Complexity of issues in moral decision-making

We now present three examples where different moral considerations lead to different courses of action. It should be emphasized that we are focusing on moral issues, and not on legal or other such considerations.

Case of the Muslim boy who almost joined the Islamic State (Goldman 2015): The 19-year-old son of a Muslim American family was at the Istanbul airport on his way to Syria for joining the Islamic State. His family persuaded him to turn back, and he returned to their home in the Houston (Texas) suburbs. One-and-a-half years later, he was charged with conspiracy and attempting to provide material support to the Islamic State. From an ethical point of view, there are reasons to put him behind bars, as he poses a threat to the society. He once came close to the edge, and though he withdrew that time, there is no surety that he will not cross over again, or will not express his allegiance to the Islamic State in some other way by harming people around him. On the other hand, if he is sent to the prison, it will deter other recruits who are planning to cross over from turning back, for if they do so they might be facing potentially long prison sentences.

Though this issue seems to have arisen in the context of the current atmosphere of terrorism and the turmoil in the Middle East, including the rise of Islamic State, the situation was very similar in the United States about a hundred years ago, when the society was feeling threatened by anarchists and communists. Healy (2013) shows how Oliver Wendell Holmes changed his views from the "sacred right to kill the other fellow when he disagrees" to his famous dissent in *Abrams v. United States,* where he forcefully argued that the First Amendment is there to promote 'free trade in ideas', and there has to be 'clear and immediate danger' before any ideas can be suppressed. The painstakingly detailed historical research by Healy demonstrates how the interplay of personal experiences, discussion with friends and colleagues, opinions of other philosophers and jurists etc.was instrumental in the evolution of Holmes's opinion. We would like to be able to model this evolution.

Crash of Germanwings Flight 9525: On 24 March 2015, the co-pilot of this flight from Barcelona to Dusseldorf locked the pilot out of the cockpit and crashed the plane in the French Alps, killing all the 144 passengers and six crew members. Later enquiry revealed that the co-pilot had been treated for severe depression and suicidal tendencies and has been declared unfit to work, but he hid this information from his employer.

In the light of this tragedy, what action should we take? Should we force therapists to provide information on their client's mental state to their employer, if they feel that the client is going to snap? Should we ban anyone who has been treated for depression from taking on tasks like an airplane pilot? (Some of these issues are addressed in Shpancer 2015.)

As in the previous examples, there have been similar cases in the past. For example, US District Court Judge Martin Feldman presided over the case of the blanket moratorium on all deep-water offshore oil-drilling put in place by the US Government in the wake of a catastrophic explosion in DeepWater Horizon, a deep-water oil rig in the Gulf of Mexico. Environmentalists argued this tragedy illustrates that deep-water oil drilling is a risky business, and the environmental cost of continuing this is too high. Judge Feldman, however, disagreed, and wrote in his decision: "[N]ot all trains are dangerous. Not all planes are going to crash. I looked at the tragedy of the DeepWater Horizon as a horrible incident in an industry in which, statistically, it was immensely rare for something like that to happen. And that's how I approached the case." (Cohen 2014).

Later on, justifying his decision, Judge Feldman compared this situation to the Boston Marathon bombing in April 2013. It would not be right, he argued, to ban all marathons in the wake of Boston Marathon's bombing unless some specific evidence indicates an impending threat to other marathons. Our goal is to have a system that can generate such arguments that are persuasive for humans.

*Case of sex robots:*In recent years, technology has made it possible to realize full-body active and intelligent sex robots such as imagined in the movie *Blade Runner* and,more recently, in *Ex Machina.* This is raising a number of ethical issues. Some argue that such robots should be banned because they reinforce the stereotype of women as sex objects. Others argue that such robots allow technology to offer people happiness and fulfillment.[1] Some of these arguments are related to the issues raised for and against legalizing prostitution (Trifiolis 2014). For instance, given

[1]http://www.bbc.com/news/technology-34118482. Accessed on 22 Sept. 2015. See also Levy (2008).

that these robots are designed to be submissive, does it encourage the expectation of a similar behavior from a real woman in the user of such a robot?

There are other issues as well, with legal consequences. For example, if someone buys such a sex robot, and spends more and more time with it, does this count as infidelity and constitutes grounds for a divorce?[2]In this regard, previous cases involving affairs through virtual reality (like in Second Life)[3], and through chatrooms (Ben-Zeév 2008) become relevant.Another issue is, if someone gets addicted to a sex robot, does the manufacturer of the sex robot bears the rsponsibility? Can the user of the sex robot sue the maker of the sex robot on grounds similar to the ones used to sue the tobacco companies, or fast-food chains? We expect our systems to be able to come up with such arguments with supporting rationales.

Modeling human decision-making

Human decision-making is a complex process and diverse paradigms have been defined to model it. The ASPECT model (Jameson et al. 2014) distinguishes a set of six choice patterns, namely, attribute-based, social-based, consequence-based, experience-based, policy-based and trial-and-error-based, which may be present separately or in combinations. Such patterns are often used to model the selection process in the recommendation systems, but they are also present in the moral decision-making task.

There aredifferent approaches to ethical reasoning that model distinct aspects of human choices, and simulating moral decision-making requires integrating these diverseapproaches (Dehghani, et al. 2008). For example, the MoralDMsystem (Blass and Forbus 2015) combines the rule-based and the analogical reasoning modules to consider both the utilitarian and deontological approaches. In our framework, we aim to incorporate experts representing different elements of the ASPECT model to analyze the situation from multiple points of view.

Jameson et al. (2014) also noted that there are two distinct approaches to designing choice architectures: they may either *persuade* (by introducing bias) or *support choice* (by presenting unbiased possibilities).In our design, we aim to *support* the user by providing set of possible action choices with moral justification for each possibility.

The problem of moral decision-making may be seen as similar to that of generating recommendations automatically (Ricci et al. 2010). In a recommender system, the goal is to find the most appropriate item according users' preferences, or some other constraints and social influences. For our case, the recommended items are the moral decisions to be taken, and the system is aimed to support the user in the selection process. As observed in (Jameson et al. 2014), a good recommender system not only provides recommendationsthat the user is likely to like, but also gives the user rationales behind the choices offered. It is interesting to note that one of the first expert systems*Mycin* (Shortliffe 1976) was found to be lacking in explanations, and this feature was added later in *Emycin* (Ulug 1986). We are currently using a hierarchical multi-agent architecture for implementing a recommender system that generates a rationale for each of the choices recommended to the user (Misztal& Indurkhya 2015). This feature is even more critical for a moral decision-making system, where the reasons accompanying the choices may be crucial in determining which action is actually taken by the user.

A system for generating moral arguments

We plan to use a multi-agent architecture known as the blackboard model (Carver & Lesser 1992; Nii 1986) to implement a system for generating moral arguments. The blackboard model is often used to represent complex and ill-defined problems that require analysis from diverse points of view. In this architecture, a group of independent experts representing diverse knowledge sources interact using a common workspace (the *blackboard)* where all the information about the problem as well the partial solutions is stored. The blackboard system allows combining diverse sources of knowledge such as rule-based as well as precedence-based modules. It also provides a hierarchical structure with multiple levels of abstractions over which top-down and bottom-up processes act in consort to generate an argument (or a diagnosis). In our earlier research, we have successfully incorporated the blackboard architecture in a poetry generating system (Misztal & Indurkhya 2016), and now we plan to adapt this architecture to implement a support system for moral decision-making.

Proposed System Architecture

Our proposed system architecture for a support system for moral decision-making is shown in Fig. 1. Its major components are described below.

Blackboard is a common workspace where various experts interact and develop solutions (rationales for moral decisions). It has multiple layers arranged in a hierarchy: at the lowest level are the concrete facts, and the highest level contains moral justifications; intermediate layers have concepts at different levels of abstractions, though only one such layer is shown Fig. 1.

[2] See also https://www.youtube.com/watch?v=2MeQcI77dTQ. Accessed on 15 Sept. 2015.

[3] http://www.theguardian.com/technology/2008/nov/13/second-life-divorce. Accessed on 22 Sept. 2015.

Initially, the details of the given situation, which is being considered from a moral point of view, are placed on the blackboard. These may be at the level of concrete details, and/or at any of the concept levels. For example, in the case of the Muslim boy, the details of the case are inputted at the concrete level, and possible actions (putting the boy in the prison, for example) are put at an appropriate concept level. For the case of sex robots, only possible actions (put a ban on sex robots) are put at a concept level.

As the process of generating different moral arguments for or against taking any particular action proceeds, the blackboard contains various partial arguments. Parts of these arguments at different hierarchical layers are interconnected to reflect abstraction hierarchy among concepts.

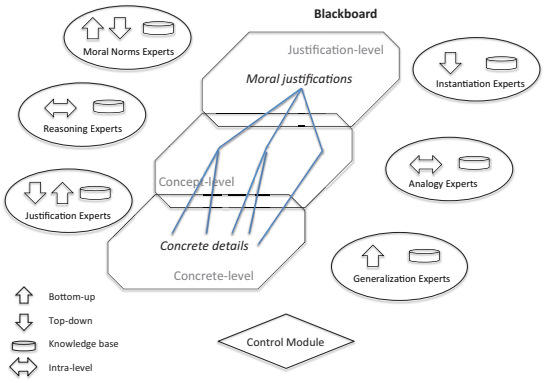

Figure 1: *Blackboard architecture for a moral decision-support system.*

Experts are independent modules representing distinct knowledge sources that have access to thecommon blackboard. They are triggered by events on the blackboard; when an expert is activated, it processes some piece of existing information on the blackboard, and posts the resulting information on the blackboard. The triggering information and the posted information may be at the same or different layers of the blackboard.

The experts can be grouped according to whether they work within the same layer or across multiple hierarchical layers, or according to the kind of expertise they incorporate. According to the layer-based grouping, we get three kinds of experts:

Top-down experts work as follows: An expert, embodying the rule "if someone is a threat to the society then that person may be imprisoned," may post a query at a lower layer: "determine if a person who intended to join IS, but changed his mind at the last minute, is a threat to the society." Other experts will then try to provide supporting or refuting evidence for this query.

Intra-level experts connect facts and make inferences at the same hierarchical level. For instance, in the example of Germanwings crash, an intra-level expert may infer that if "require therapists to disclose information about severely mentally ill patients to their employers," then "mentally ill patients may be reluctant to go to a therapist." Another intra-level agent may infer that given the facts that a pilot who was suffering from severe depression and suicidal thoughts, and was declared unfit to work, still flew as a co-pilot and deliberately crashed the plane killing all the passengers, it would be advisable to prevent a pilot with mental illness to continue flying planes. Yet another intra-level rule might infer that if a therapist deems a pilot mentally unfit for flying, they should be required to notify this to the authorities.

Bottom-up experts work as follows: Given the suggestion that if a therapist deems a pilot mentally unfit for flying, they should be required to notify this to the authorities, a bottom-up expert may post a rule on a higher layer: "require therapists to disclose information about severely mentally ill patients who drive passenger-carrying vehiclesto their employers".

Grouped by the kind of expertise they embody, we get the following kinds of experts:

Moral norms experts: These embody moral principles like, 'things should be fair', or 'the society should be protected from harm by individuals.' They can be either top-down or bottom-up: information at a lower level can trigger a moral-norm expert, and a moral-norm expert, when activated, can post a task at a lower layer.

Generalization experts: These are bottom-up experts,which, based on the information posted at a layer, post a generalized statement or concept at a higher layer. For example, 'sex with robots' may be generalized to 'technology-mediated relationship; or 'airplane pilot' can be generalized to 'operator of passenger-carrying vehicles'.

Instantiation experts: These top-down experts work in the opposite direction to the generalization experts by instantiating a more concrete instance in a lower layer based on some more general statement in a higher layer. For example, 'technology-mediated relationships' may create instances like 'chat-room relationship' or 'second life relationships'.

Analogy experts: Instantiation and generalization experts together can create analogous siblings within a layer. Then intra-level analogy experts explore similarities and differences between these siblings to figure out if moral justification for one can be applied to another.

Reasoning experts: These intra-level experts infer new information based on the information posted in a layer.

Justification experts: These experts create justifications for different statements or beliefs on the blackboard. They can be either top-down or bottom-up. A top-down justification expert works by posting a query or seeking

some information at a lower level to justify a belief or statement. A bottom-up justification expert works by posting a belief or statement at a higher layer based on the information contained in a lower layer.

Controller is a module that determines the order in which the activated experts are executed.

Conclusions and future research

We proposed here a multi-agent blackboard architecture for supporting moral-decision making, and presented an example of possible application of the system for some real-life moral dilemmas. The model is flexible and enables incorporating multiple independent modules that represent diverse sources of knowledge, and therefore has a potential to model cognitive processes. We compared our research to the current state of the art in the ethical choices supporting systems and grounded it on the models of human decision-making behavior. Further work will focus on the implementation and evaluation of the system by testing it with real-life scenarios.

References

Ariely, D. 2010. Predictable irrational: The hidden forces that shape our decision. Harper Perennial.

Ariely, D. 2013. The Honest Truth About Dishonesty: How We Lie to Everyone--Especially Ourselves. Harper Perennial.

Arkin, R.C., Ulam, P. & Wagner, A.R. 2012. Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust and deception. Proceedings of the IEEE 100(3).

Ben-Zeév, A. 2008. Is chatting cheating? Psychology Today(Sept. 05, 2008).Accessed on 22 Sept. 2015. https://www.psychologytoday.com/blog/in-the-name-love/200809/is-chatting-cheating.

Blass, J.A. &Forbus, K.D. 2015. Moral decision-making by analogy: Generalizations versus exemplars. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.

Bucciarelli, M., Khemlani, S. &Johnson-Laird, P.N. 2008. The psychology of moral reasoning.Judgment and Decision Making3(2), 121–39.

Carver, N. &Lesser, V. 1992. The evolution of blackboard control architectures.Technical report, CMPISCI.

Churchland, P. 2012. Braintrust: What neuroscience tells about morality.Princeton University Press.

Cohen, J. 2014. Blindfolds Off: Judges on How They Decide. American Bar Association.

Damasio, A. 2005. Descartes' error: Emotion, reason and the human brain.Penguin Books (reprint edition).

Dehghani, M.; Tomai, E.; Forbus, K. and Klenk, M. 2008. An integrated reasoning approach to moral decision-making. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence 1280-1286.

Goldman, A. 2015. An American family saved their son from joining the Islamic State. Now he might go to prison. The Washington Post (Sept. 6, 2015)

Healy, T. 2013 The Great Dissent: How Oliver Wendell Holmes changed his mind and changed the history of free speech in America. New York: Metropolitan Books.

Herman, G. 2011. Moral reasoning. Accessed on 21 Sept. 2015. http://www.princeton.edu/~harman/Papers/Moral_Reasoning_Current.pdf.

Herman, G. 2015. Moral relativism explained. http://www.princeton.edu/~harman/Papers/Moral%20Relativism%20Explained.pdf. Accessed on 21 Sept. 2015.

Indurkhya, B. 2015. A cognitive perspective on norms.To appear in B. Brożek (ed.) The Normative Mind. Cracow (Poland): The Copernicus Center Press.

Indurkhya, B. & Misztal, J. 2015. On modeling cognitive and affective factors in legal decision-making. In A. Rotolo (ed.). Legal Knowledge and Information Systems, Amsterdam: IOS Press, 157-160.

Jameson A., Berendt B., Gabrielli S., Gena C., Cena F., Vernero F., and Reinecke K. 2014 Choice Architecture for Human-Computer InteractionFoundations and Trends in Human-Computer Interaction 7(1–2), 1–235.

Johnson, M. 2014. Morality for Humans. University of Chicago.

Kant, I. 1785. Groundwork of the metaphysics of morals, New York: Cambridge University Press (2nd ed.: 2012).

LeDoux, J. 2003. The synaptic self: How our brains become who we are. Penguin Books.

Levy, D. 2008. Love and sex with robots: The evolution of human-robot relationships.

Lin, P., Abney, K. &Bekey G.A. (eds.) 2014.Robot Ethics: The Ethical and Social Implications of Robotics. The MIT Press.

Misztal, J. & Indurkhya, B.2016. A blackboard system for generating poetry.To appear in Computer Science Journal.

Misztal, J. &Indurkhya, B. 2015. Explaining contextual recommendations: Interaction design study and prototype implementation. Proceedings of the Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (9th ACM Conference on Recommender Systems), Vienna (Austria).

Nii, H. 1986. Blackboard systems: The blackboard model of problem solving and the evolution of blackboard architectures. AI Magazine, 7.

Ricci, F., Rokach, L.,Shapira, B. and Kantor, P. B. 2010Recommender Systems HandbookSpringer-Verlag New York, Inc.

Richtel, M. &Dougherty, C. 2015.Google's Driverless Cars Run Into Problem: Cars With Drivers. The New York Times(Sept. 1, 2015).

Shortliffe, E.H. 1976. Computer-based medical consultations: MYCIN.New York: Elsevier/North Holland.

Shpancer, N. 2015. How to understand the Germanwings crash: Finding potential disasters before they strike is harder than it seems. Psychology Today (May 4, 2015). Accessed on 15 Sept. 2015.https://www.psychologytoday.com/blog/insight-therapy/201505/how-understand-the-germanwings-crash.

Trifiolis, K. 2014. Legalizing Prostitution: An Introduction. Student Scholarship, paper 139. Accessed on 23 Sept. 2015. http://erepository.law.shu.edu/student_scholarship/139.

Ulug, F. 1986.Emycin-Prolog expert system shell.Master's Thesis. Naval Postgraduate School, Monterey, California.